Universität Augsburg

Institut für Mathematik

Antony Unwin

Good Graphics?

Preprint Nr. 027/2007 — 11. Juli 2007

Institut für Mathematik, Universitätsstraße, D-86 135 Augsburg http://www.math.uni-augsburg.de/

Impressum:

Herausgeber:

Institut für Mathematik
Universität Augsburg
86135 Augsburg
http://www.math.uni-augsburg.de/forschung/preprint/

ViSdP:

Antony Unwin
Institut für Mathematik
Universität Augsburg
86135 Augsburg

Good Graphics?

Antony Unwin

Augsburg University, Germany unwin@math.uni-augsburg.de

Graphical Excellence is nearly always multivariate.

Edward Tufte

1 Introduction

This chapter discusses drawing good graphics to visualize the information in data. Graphics have been used for a long time to present data. Figure 1 is a scanned image from Playfair's *Commercial and Political Atlas* of 1801, reproduced in Playfair (2005). The fairly continuous increase of both imports and exports, and the fact that the balance was in favour of England from 1720

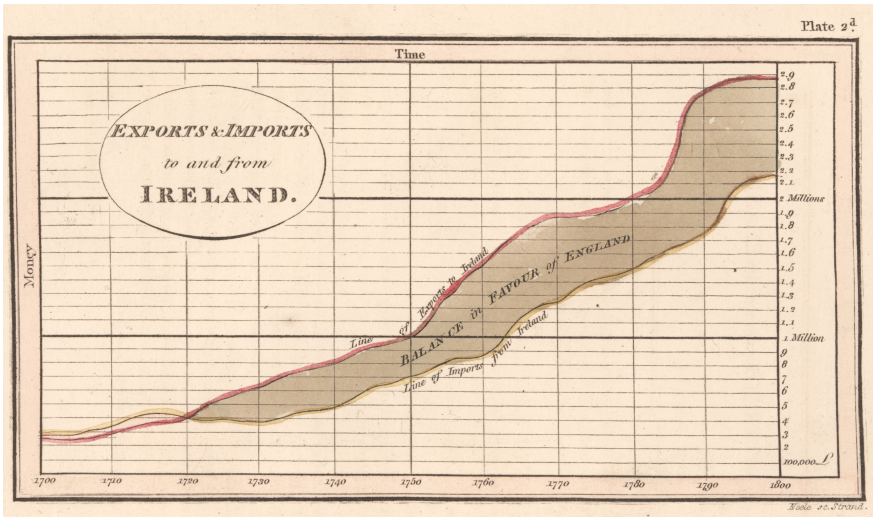

Fig. 1. Playfair's chart of trade between England and Ireland from 1700 to 1800.

on, can be seen easily. Some improvements might be made, but overall the display is effective and well drawn.

Data graphics are used extensively in scientific publications, in newspapers, and in the media generally. Many of those graphics do not fully convey the information in the data they are supposed to be presenting and may even obscure it. What makes a graphic display of data bad but, more importantly, what makes one good? In any successful graphic there must be an effective blending of content, context, construction and design.

1.1 Content, Context, and Construction

What is plotted comes first, and without content no amount of clever design can bring meaning to a display. A good graphic will convey information, but a graphic is always part of a larger whole, the context, which provides its relevance. So a good graphic will complement other related material and fit in, both in terms of content and also with respect to style and layout. Finally, if a graphic is constructed and drawn well, it will look good.

Figure 2 shows two similar displays of the same data from the DDB social survey used in Robert Putnam's book *Bowling Alone* (Putnam; 2000). Every year for twenty-four years, different groups of 3000 people were surveyed. Amongst other questions they were asked how often they had attended church in the last year.

The left-hand graph includes gridlines and a coloured background and uses three-dimensional columns to represent the data counts. The right-hand graph sticks to basics. In general, the right-hand display is to be preferred (three-dimensional columns can cross gridlines and zero values would be misleadingly displayed). For these data there is not much to choose between the two representations, both convey the same overall information. The potential weakness in both graphics is the set of categories. Grouping the data together in different ways could give quite different impressions.

For a given dataset there is not a great deal of advice which can be given on content and context. Those who know their own data should know best

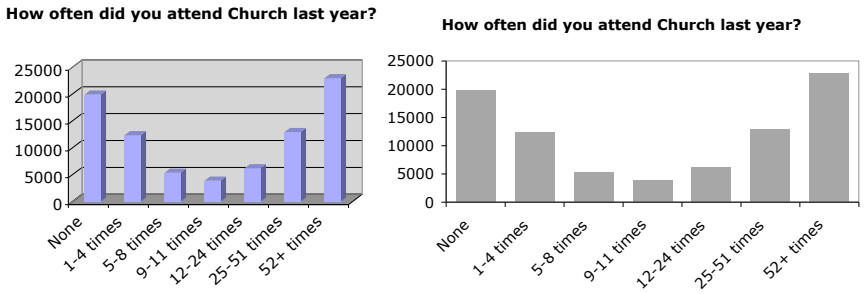

Fig. 2. Church attendance (DDB Life Style Survey 1975-1998)

for their specific purposes. It is advisable to think hard about what should be shown and to check with others if the graphic makes the desired impression. Design should be left to designers, though some basic guidelines should be followed: consistency is important (sets of graphics should be in similar style and use equivalent scaling); proximity is helpful (place graphics on the same page, or on the facing page, of any text that refers to them); and layout should be checked (graphics should be neither too small nor too large, and be attractively positioned relative to the whole page or display). Neither content nor context nor design receive much attention in books offering advice on data graphics, quite properly they concentrate on construction. This chapter will too.

1.2 Presentation Graphics and Exploratory Graphics

There are two main reasons for using graphic displays of datasets, either to present data or to explore data. Presenting data involves deciding what information you want to convey and drawing a display appropriate for the content and for the intended audience. You have to think about how the plot might be perceived and whether it will be understood as you wish. Plots which are common in one kind of publication may be unfamiliar to the readers of another. There may only be space for one plot and it may be available in print for a very long time, so great care should be taken in preparing the most appropriate display. Exploring data is a much more individual matter, using graphics to find information and to generate ideas. Many displays may be drawn. They can be changed at will or discarded, and new versions prepared, so generally no one plot is specially important, and they all have a short life span. Clearly principles and guidelines for good presentation graphics have a role to play in exploratory graphics but personal taste and individual working style also play important roles. The same data may be presented in many alternative ways, and taste and customs differ as to what is regarded as a good presentation graphic. Nevertheless, there are principles that should be respected, and guidelines that are generally worth following. No one should expect a perfect concensus where graphics are concerned.

2 Background

2.1 History

Data graphics may be found going very far back in history but nowadays most agree that they really began with the work of Playfair a little more than two hundred years ago. He introduced some modern basic plots (including the barchart and the histogram) and he produced pertinent and eye-catching displays (see Figure 1). Wainer and Spence have recently republished a collection of his works (Playfair; 2005). Not all his graphics could be described as good,

but most were. In the second half of the nineteenth century Minard prepared impressive graphics, including his famous chart of Napoleon's advance on and retreat from Moscow. The French Ministry of Public Works used his ideas to attractive, and presumably pertinent, effect in an annual series of publications (*Album de Statistique Graphique*) from 1879-1899, presenting economic data geographically for France. Examples can be found in Michael Friendly's chapter in this book.

In the first half of the last century graphics were not used in statistics as much as they might have been. Interestingly, the second chapter in Fisher's *Statistical Methods for Research Workers* in 1925 was on diagrams for data, so he, at least, thought graphics important. In Vienna there was a group led by Otto Neurath, which worked extensively on pictograms in the 1920s and early 1930s. They produced some well-crafted displays, which were forerunners of the modern Infographics. (Whether Figure 3 is improved by including the symbols at the top to represent the USA is a matter of taste.)

With the advent of computers, graphics went into a relative decline. Computers were initially bad for graphics for two reasons. Firstly, because much more complex analytic models could be evaluated and, quite naturally, modelling received a great deal more attention than displaying data. Secondly, because only simple and rather ugly graphics could be drawn by early computers. The development of hardware and software has turned all this around. In recent years it has been very easy to produce graphics, and far more are to be seen than before. Which is, of course, all the more reason to be concerned that the graphics are drawn well.

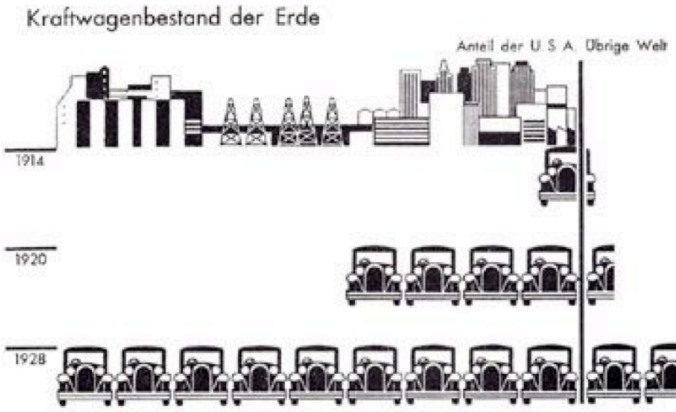

Fig. 3. A pictogram by Otto Neurath of the number of cars in the US and the rest of the world in 1914, 1920 and 1928.

2.2 Literature

Several authors have written excellent books on drawing good statistical graphics, the best known, justifiably, being Edward Tufte. His books, e.g., (Tufte; 2001), include many splendid examples (and a few dreadful ones), and describe important principles on how to draw the good ones. Tufte criticises unsuitable decoration and data misrepresentation but his advice is restricted to representing data properly. Cleveland's books, (for instance, Cleveland (1994)), another useful source of advice on preparing data displays, are the same. And this is the way it should be. Statisticians should concentrate on getting the basic statistical display right, designers may be consulted to produce a polished final version.

While there is a place for applied books full of sound practical advice, (other useful references include Burn (1993), Kosslyn (1994) and Robbins (2004)), there is also a need for theory to provide formal structures for understanding practice and to provide a foundation from which progress can be made. Graphics must be one of the few areas in statistics where there is little such theory. Bertin's major work (*Semiologie Graphique* Bertin (1973)) contains a number of interesting ideas and is often cited, but it is difficult to point to later work that directly extends it. Wilkinson's *Grammar of Graphics* has received a lot of attention and been quickly revised in a substantially expanded second edition (Wilkinson; 2005).

If there is little theory, then examples become particularly important to show what can be achieved. The two books by Wainer, (Wainer; 1997) and (Wainer; 2004), contain collections of columns first published in Chance and offer instructive and entertaining examples. Friendly's Gallery of Statistical Visualization (http://www.math.yorku.ca/SCS/Gallery/) includes many examples, both good and bad, chronicling the history of graphical developments. The websites ASK E.T. (http://www.edwardtufte.com) and Junk Charts (http://junkcharts.typepad.com) provide lively discussions and sage advice for particular examples. It would be invidious, and perhaps unfair, to single out egregious examples here. Readers should be able to find plenty for themselves without having to look far.

2.3 The Media and Graphics

Graphical displays of data appear in the press very often. They are a subset of Infographics, graphical displays for conveying information of any kind, usually discussed under the heading Information Visualization (Spence; 2001). Many impressive examples can be found in the New York Times. While there are guidelines which apply to all Infographics, this chapter is restricted to the construction of data visualizations.

Data displays in the media are used to present summary information, such as: the results of political surveys (what proportion of the people support which party); the development of financial measures over time (a country's trade balance or stock market indices); comparisons between population

groups (average education levels of different sections of the community). There are many other examples. These topics only require fairly basic displays, so it is not surprising that in the media they are commonly embellished with all manner of decoration and ornamentation, sometimes effectively drawing attention both to the graphic and to its subject, sometimes just making it more difficult to interpret the information being presented. What is surprising is that the graphics are often misleading or flawed.

3 Presentation (What to Whom, How and Why)

How is it possible to make a mess of presenting simple statistical information, surely there is little that can go wrong? It is astonishing just what distortion can be introduced: misleading scales may be employed; three dimensional displays of two dimensional data make it difficult to make fair comparisons; areas which are apparently intended to be proportional to values are not; so much information is crammed into a small space so that nothing can be distinguished. While those are some of the technical problems that can arise, there are additional semantic ones. A graphic may be linked to three pieces of text: its caption, a headline and an article it accompanies. Ideally, all four should be consistent and complement each other. In extreme cases all four can tell a different story! A statistician can not do much about headlines (possibly added or amended by a subeditor at the last minute) or about accompanying articles if he or she is not the first author (in the press the journalist chooses the graphic and may have little time to find something appropriate), but the caption and the graphic itself should be 'good'.

There are displays to be found in the media to highlight a news item or to provide an illustration to lighten the text. These are often prepared by independent companies at short notice and sold to the media as finished products. Fitting the graphic to its context may be awkward. There are displays in scientific publications that are prepared by the authors and should be the product of careful and thorough preparation. In this situation a graphic should match its context well. Whatever kind of data graphic is produced, there are a number of general principles to be followed to ensure that the graphic is at least correct.

Whether a graphic is then successful as a display or not depends on its subject, on its context and on aesthetic considerations. It depends on what it is supposed to show, on what form is chosen for it, and on its audience. Readers familiar with one kind of graphic will have no trouble interpreting another example of the same kind. On the other hand, a graphic in a form which is new to them may lead to unanticipated interpretation difficulties. When someone has spent a long time on a study and further time on the careful prepartion of a graphic display to illustrate the conclusions, they are usually astonished when others do not see what they can see. (This effect is, of course, not restricted to drawing graphics. Designers are frequently shocked

by how people initially misunderstand their products. How often have you stared at the shower in a strange hotel wondering how you can get it to work without its scalding or freezing you? Donald Norman's book (Norman; 1988) is filled with excellent examples.)

Other factors have to be considered as well. A graphic may look different in print than on a computer screen. Complex graphics may work successfully in scientific articles where the reader takes time to fully understand them. They will not work well as a brief item in a television news programme. On the other hand, graphics which are explained by a commentator are different from graphics in print. If graphics displayed on the web can be queried (as with some of the maps on http://www3.cancer.gov/atlasplus/, discussed in Section 5.5), then more information can be provided without cluttering the display.

4 Scientific Design Choices in Data Visualization

Plotting a single variable should be fairly easy. The type of variable will influence the type of graphic chosen. For instance, histograms or boxplots are right for continuous variables, while barcharts or piecharts are appropriate for categorical variables. In both cases other choices are possible too. Whether the data should be transformed or aggregated will depend on the distribution of the data and the goal of the graphic. Scaling and captioning should be relatively straightforward, though need to be chosen with care.

It is a different matter with multivariate graphics, where even displaying the joint distribution of two categorical variables is not simple. The main decision to be taken for a multivariate graphic is the form of display, though the choice of variables and their ordering are also important. In general a dependent variable should be plotted last. In a scatterplot it is traditional to plot the dependent variable on the vertical axis.

4.1 Choice of Graphical Form

There are barcharts, piecharts, histograms, dotplots, boxplots, scatterplots, roseplots, mosaic plots and many other kinds of data display. The choice depends on the type of data to be displayed (e.g. univariate continuous data cannot be displayed in a piechart and bivariate categorical data cannot be displayed in a boxplot) and on what is to be shown (e.g., piecharts are good for displaying shares for a small number of categories and boxplots are good for emphasising outliers). A poor choice of the type of graph cannot be rectified by other means, so it is important to get it right at the start. However, there is not always a unique optimal choice and alternatives can be equally good or good in different ways, emphasising different aspects of the same data.

Given an appropriate form has been chosen there are many options that can be considered. Simply adopting the default of whatever computer software is being used is unlikely to be wise.

4.2 Graphical Display Options

Scales

Defining the scale for the axis for a categorical variable is a matter of choosing an informative ordering. This may depend on what the categories represent or on their relative sizes. For a continuous variable it is more difficult. The endpoints, the divisions, and the tickmarks have to be chosen. Initially it is a surprise when apparently reliable software produces a really bad scale for some variable. It seems obvious what the scale should have been. It is only when you start trying to design your own algorithm for automatically determining scales that you discover how difficult the task is.

In *Grammar of Graphics* Wilkinson puts forward some plausible properties that 'nice' scales should possess and suggests a possible algorithm. The properties (simplicity, granularity and coverage, with the bonus of being called 'really nice' if zero is included) are good but the algorithm is easy to outwit. This is not to say that it is a weak algorithm. What is needed is a method which gives acceptable results for as high a percentage of the time as possible, and the user must also check the resulting scale, and be prepared to amend it for his or her data. Difficult cases for scaling algorithms arise when data cross natural boundaries, e.g., data with a range of 4 to 95 would be easy to scale, whereas data with a range of 4 to 101 would be more awkward.

There is a temptation to choose scales running from the minimum to the maximum of the data but this means that some points are right on the boundaries and may be obscured by the axes. Unless the limits are set by the meaning of the data (e.g. exam marks from 0 to 100, neither negative marks nor marks more than 100 are possible (usually!)), it is good practice to extend the scales beyond the observed limits and to use readily understandable rounded values. There is no obligatory requirement to include zero in a scale, but there should always be a reason for not doing so, otherwise it makes the reader wonder if some deception is being practised. Zero is in fact not the only possible baseline or alignment point for a scale, albeit the most common one. A sensible alignment value for ratios is one and financial series are often standardised to all start at 100. In Figure 11 the cumulative times for all the riders who finished the Tour de France cycle race in 2004 have been plotted. The data at the end of each stage have been aligned at their means. Interest lies in the differences in times between the riders, not so much in their absolute times.

Figure 4 shows histograms for the Hidalgo stamp thickness data (Izenman and Sommer; 1988). The first uses default settings and shows a skew distribution with possibly a second mode around 0.10. The second has rounded endpoints and a rounded binwidth, and shows stronger evidence for the second mode. The third is drawn so that each distinct value is in a different bin (the data were all recorded to a thousandeth of a millimeter). It suggests that the first mode is actually made up to two groups and that there may be evidence for several additional modes to the right. It also reveals that rounded

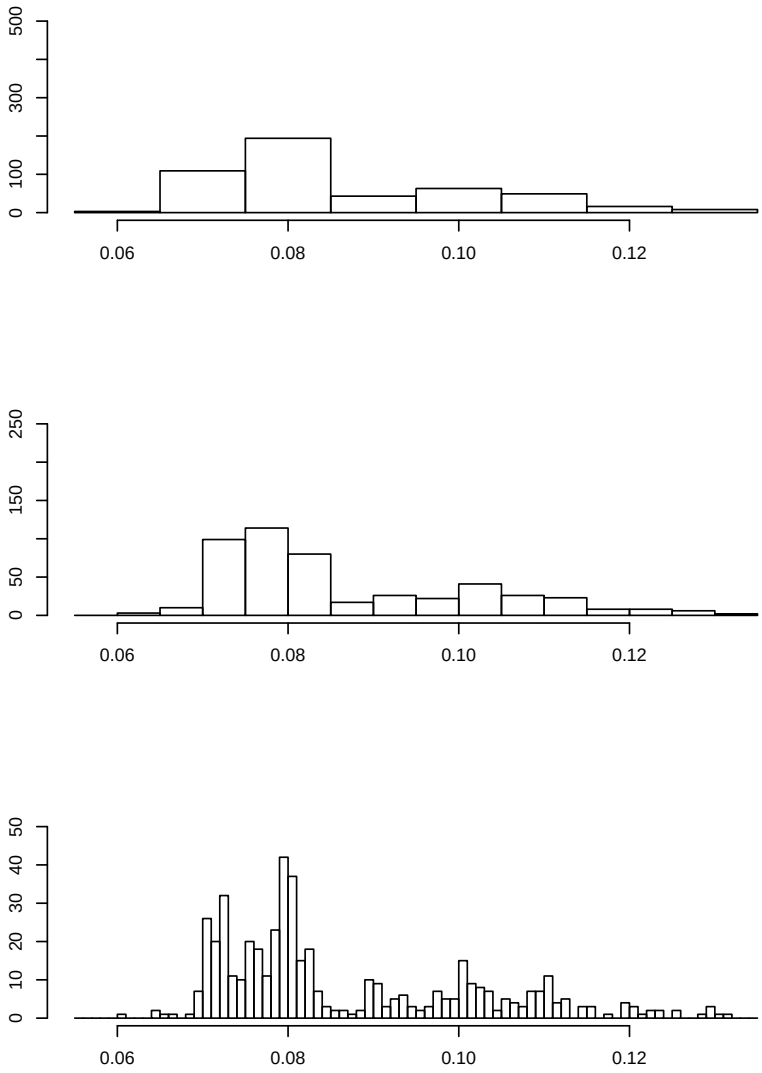

Fig. 4. Three different histograms of the Hidalgo stamp thickness data, all with the same anchorpoint but with different binwidths. The horizontal scales are aligned and the total area of each display is the same (note the different frequency scales). Source: Izenman and Sommer (1988).

values such as $0.07, 0.08, \ldots, 0.11$ occur relatively more frequently. Izenman and Sommer used the third histogram in their paper. What the data represent and how they are collected should be taken into account when choosing scales. Asymptotic optimality criteria only have a minor role to play.

While Figure 4 shows the importance of choosing binwidths carefully, it also illustrates some display issues. The horizontal value axis is clearly scaled but it would surely be nicer if it extended further to the right. More importantly, the comparison in Figure 4 ideally requires that all three plots are aligned exactly and have the same total area. Not all software provides these capabilities.

Graphics should be considered in their context. It may be better to use a scale in one graphic that is directly comparable with that in another graphic instead of individually scaling both. Common scaling is used in one form or another in Figures 11, 13, and 14.

It is one thing to determine what scale to use, but quite another to draw and label the axes. Too many labels make a cluttered impression, too few can make it hard for the reader to assess values and differences. (Note that it is not the aim of graphics to provide exact case values, tables are much better for that.) Tickmarks in between labels often look fussy and have little practical value. In some really bad situations, they can obscure data points.

Sorting and Ordering

The effect of a display can be influenced by many factors. When more than one variable is to be plotted, the position or order in which they appear in the graphic makes a difference. Examples arise with parallel coordinate plots, mosaic plots and matrix visualizations, all discussed in other chapters. Within a nominal variable with no natural ordering, the order in which the categories are plotted can have a big effect. Alphabetic ordering may be appropriate (a standard default, which is useful for comparison purposes), or a geographic or other grouping (e.g. shares by market sector) might be relevant. The categories could be ordered by size or by a secondary variable. Figure 5 shows two barcharts of the same data, the numbers in each class and in the crew on the Titanic. The second ordering would be the same in any language, but the first would vary (for instance, Crew, First, Second, Third in English).

Adding Model or Statistical Information — Overlaying (Statistical) Information

Guides may be drawn on a plot as a form of annotation and are useful for emphasising particular issues, say which values are positive or negative. Sloping guides highlight deviations from linearity. Fitted lines, for instance polynomial regression or smoothers, may be superimposed on data to show not only the hypothesised overall structure but also to highlight local variability and any lack of fit. Figure 6 plots the times from the first and last stages of a

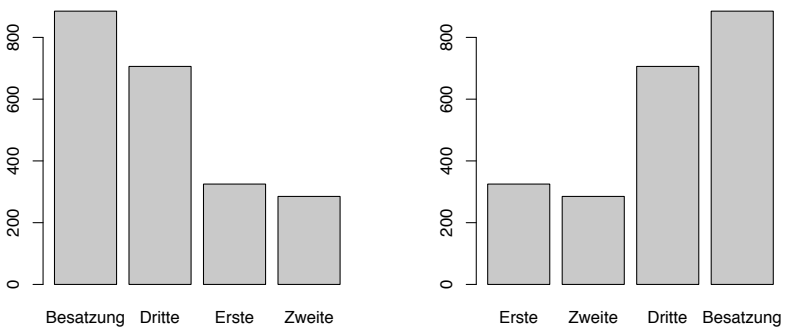

Fig. 5. Numbers of passengers and crew, who travelled on the Titanic, by class, ordered alphabetically (in German), and by status. Source: Dawson (1995).

100km road race. A lowess smoother has been drawn. It suggests that there is a linear relationship for the faster runners and a flat one for the slower ones.

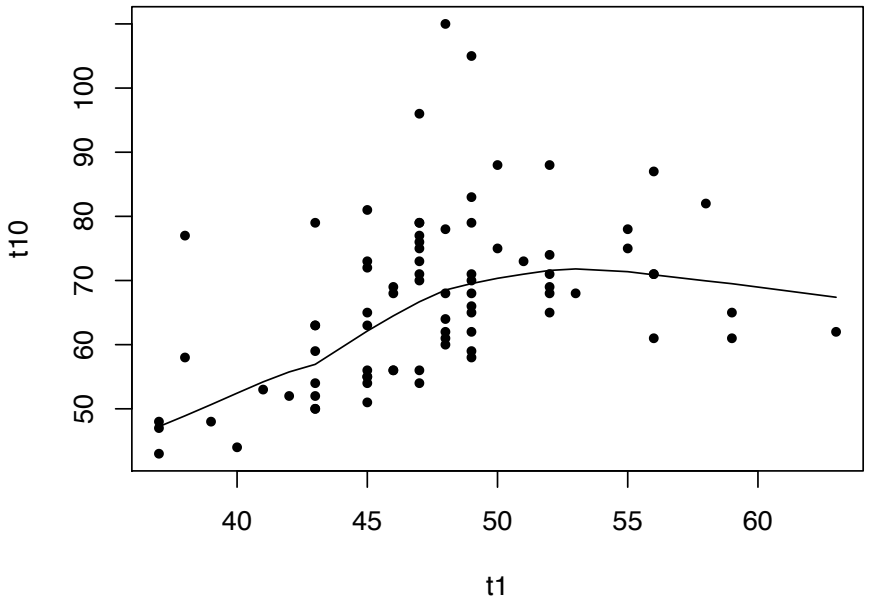

Fig. 6. Times for 80 runners for the last stage of a road race v. their times for the first stage, with a lowess smoother. Default scales from R have been used. Source: Everitt (1993).

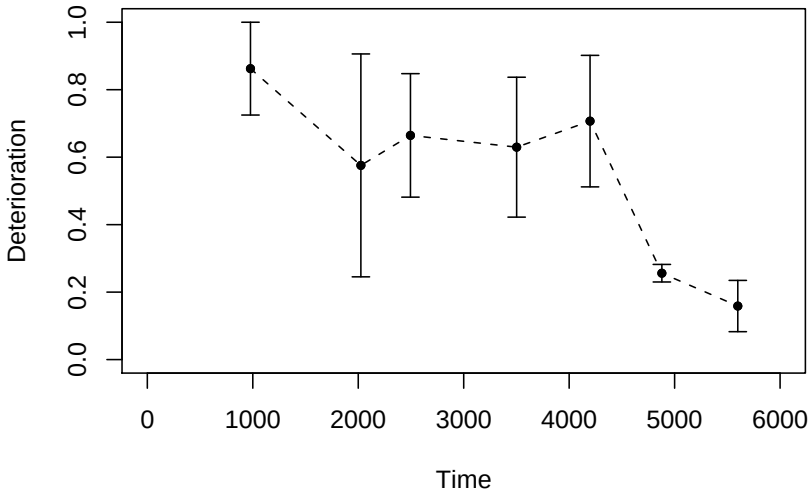

Fig. 7. Average chemical deterioration and 95% confidence intervals, measured at different time points. There were either 5 or 10 measurements at each time point. Source: Confidential research data.

When multiple measurements are available, it is standard to plot point estimates with their corresponding confidence intervals in scientific journals. (95% confidence intervals are most common, though it is wise to check precisely what has been plotted.) Figure 7 displays the results of a study on the deterioration of a thin plastic over time. Measurements could only be made by destructive testing, so all measurements are of independent samples. The high variability at most of the time points is surprising. Adjacent means have been joined by straight lines. A smoothing function would be a better alternative, but is not common for this kind of plot. As the measurement timepoints are far apart and as there is only one dataset, there is no overlapping here. That can very often be a serious problem.

Overlaying information, whether guides or annotation, can lead to overlapping and cluttered displays. Good solutions are possible but may require individual adjustments depending on the shape of the data. A well-spaced and informative display at one size may appear unsatisfactory and unclear when shrunk for publication.

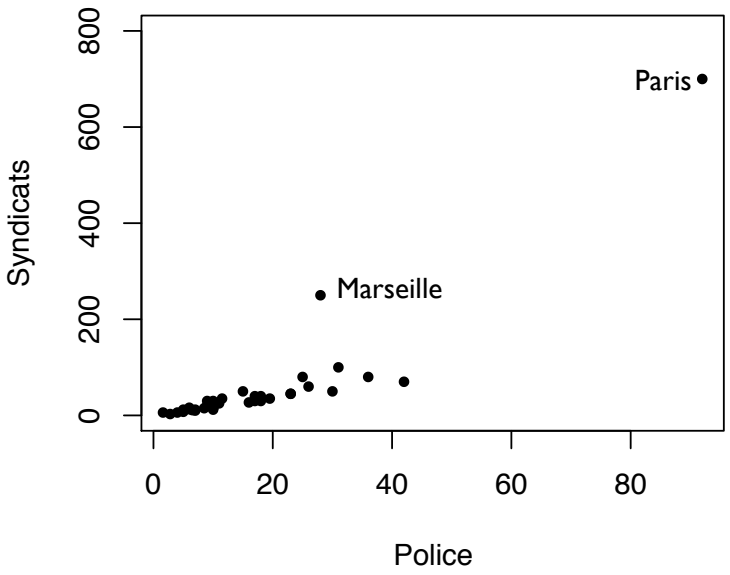

Fig. 8. Union estimates (in thousands) of the protest turnout in various French cities in the Spring of 2006 plotted against police estimates (also in thousands). Source: Le Monde 28.3.06.

Captions, Legends and Annotations

Ideally, captions should fully explain the graphic they accompany, including giving the source for the data. Relying on explanations in the surrounding text rarely works. Ideals cannot always be met and very long captions are likely to put off the reader — the whole point of a graphic is to present information concisely and directly. A compromise where the caption outlines the information in the graphic and a more detailed description can be found in the text can be a pragmatic solution. Graphics which require extensive commentary may be trying to present too much information at one go.

Legends describe which symbols and/or colours refer to which data groups. Tufte recommends that this information should be directly on the plot and not in a separate legend, so that the reader's eyes do not have to jump backwards and forwards. If it can be done, it should be.

Annotations are used to highlight particular features of a graphic. For reasons of space there cannot be many of them and they should be used sparingly. They are useful for identifying events in time series, as Playfair did, (Playfair; 2005), or for drawing attention to particular points in scatterplots.

Union estimates of protest turnout in Figure 8 are larger than the police estimates by roughly the same factor, except for the two extreme exceptions, Marseille and Paris, where the disagreement is much greater.

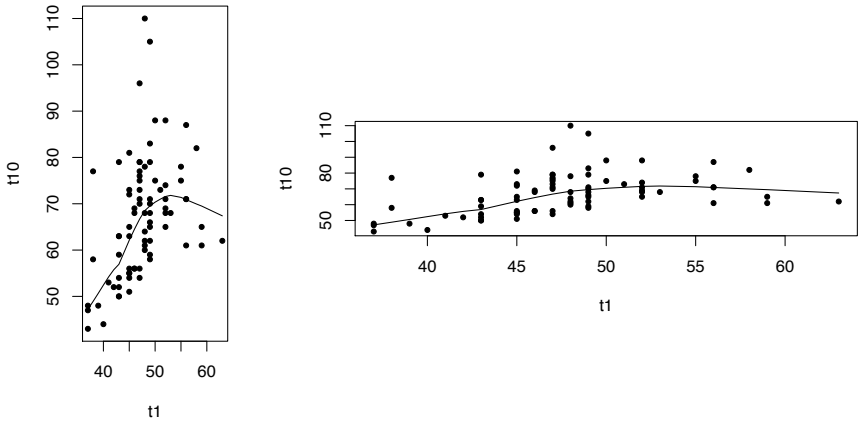

Fig. 9. The same plot as Figure 6, drawn with different aspect ratios. Data are times for runners for the last stage of a road race v. their times for the first stage.

Positioning in Text

Keeping graphics and text on the same page or on facing pages is valuable for practical reasons. It is uncomfortable to have to turn pages backwards and forwards, because graphics and the text relating to them are on different pages. However, it is not always possible. Where graphics are placed on a given page is a design issue.

Size, Frames and Aspect Ratio

Graphics should be large enough for the reader to see the information in them clearly and not much larger. This is a rough guideline, as much will depend on the surrounding layout. Frames may be drawn to surround graphics. As frames take up space and add to the clutter, they should best only be used for purposes of separation, i.e., separating the graphic from other graphics or from the text.

Aspect ratios have a surprisingly strong effect on the perception of graphics. This is especially true of time series. If you want to show gradual change, grow the horizontal axis and shrink the vertical axis. The opposite actions will demonstrate dramatic change. For a scatterplot example, see Figure 9, which displays the same data as Figure 6. There is useful advice on aspect ratios in Cleveland (1994), especially the idea of "banking to 45 degrees" for straight lines.

Colour

Colour should really have been discussed much earlier. It is potentially one of the most effective ways of displaying data. In practice it is also one of

the most difficult to get right. A helpful check for colour schemes for maps, Colorbrewer by Cynthia Brewer, can be found at http://colorbrewer.org. Colorbrewer can give suggestions for colour schemes that both blend well and distinguish between different categories.

There remain many factors in the choice of colour, which have to be born in mind: some people are colour blind; colours have particular associations (red for danger or for losses); colours may not be reproduced in print the way they were intended; and colour can be a matter of personal taste. Colour is discussed in more detail in other Handbook chapters.

5 Higher Dimensional Displays and Special Structures

5.1 Scatterplot Matrices (Sploms)

Plotting each continuous variable against every other one is effective for small numbers of variables, giving a overview of possible bivariate results. Figure 10

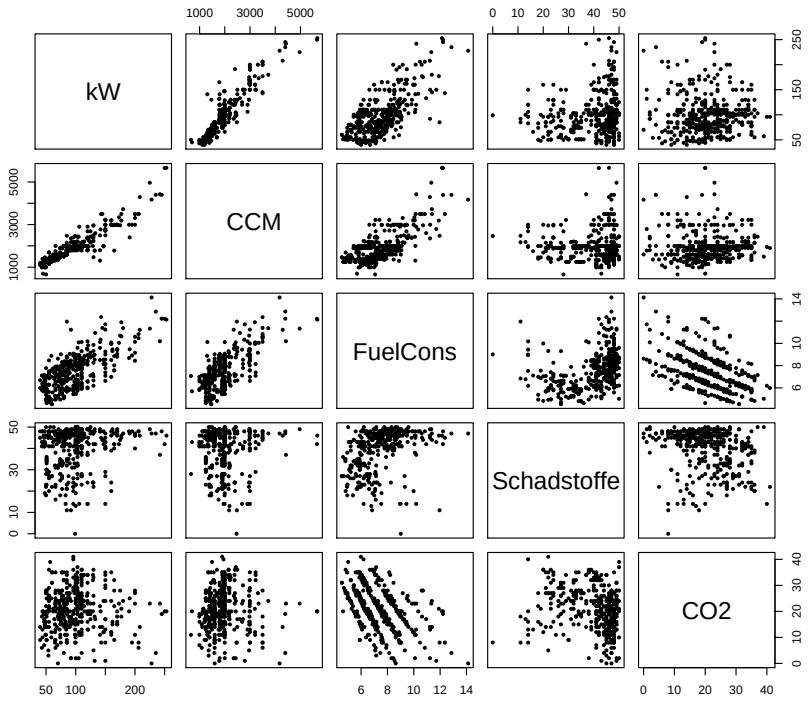

Fig. 10. A scatterplot matrix of the five main continuous variables from a car emissions dataset from Germany. Source: http://www.adac.de, March 2006.

displays data from emissions tests of 381 cars on sale in Germany. It reveals that engine size, performance and fuel consumption are approximately linearly related, as might be expected, that CO2 measurements and fuel consumption are negatively correlated in batches, which might not be so expected, and that other associations are less conclusive. Packing so many plots into a small space it is important to cut down on scales. Placing the variable names on the diagonal works well, and histograms of the individual variables could also be placed there.

5.2 Parallel Coordinates

Parallel coordinate plots (Inselberg; 1999) are valuable for displaying large numbers of continuous variables simultaneously. Showing so much information at once has several implications: not all information will be visible in any one plot (so that several may be needed); formatting and scaling will have a big influence on what can be seen (so that there are many choices to be made); and some overlapping is inevitable (so that α-blending or more sophisticated density estimation methods are useful).

Figure 11 plots the cumulative times of the 147 cyclists at the ends of the 21 stages of the 2004 Tour de France. The axes all have the same scale, so that differences are comparable. The best riders take the shortest time and are at the bottom of the plot. The axes have been aligned at their means, as without some kind of alignment little could be seen. α-blending has been applied to

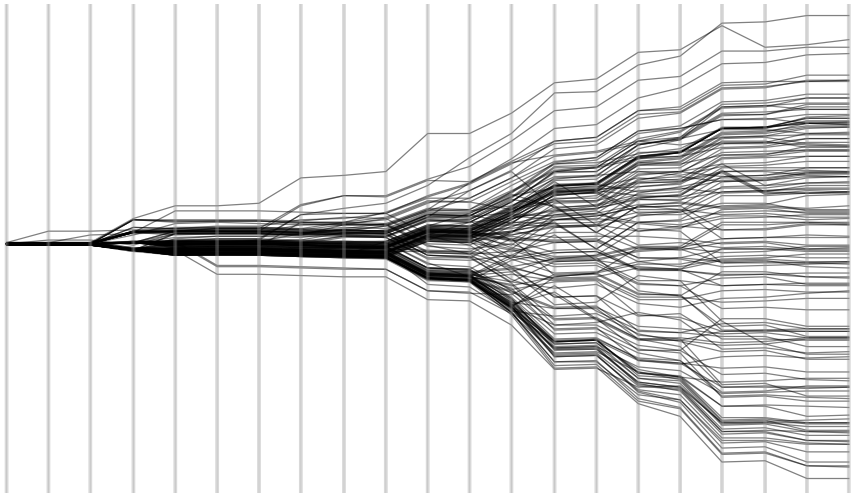

Fig. 11. Cumulative times for the riders in the 2004 Tour de France for the 21 stages. The axes have a common scale and are aligned by their means. Each vertical line represents a stage and they have been plotted in date order. Source: http://www.letour.fr.

reduce the overprinting in the early sprint stages where all riders had almost the same times. If more α-blending is used then the individual lines for the riders in the later stages of the race become too faint. This single display conveys a great deal about the race. In the early stages at most a few minutes separates the riders. On the mountain stages there are much larger differences and individual riders gain both time and places (where a line segment crosses many others downwards). Note that there are relatively few line crossings over the later stages of the race, which means, perhaps surprisingly, that not many riders changed their race ranking.

This graphic might be improved in a number of ways: the axes could be labelled (though there is little space for this); the vertical axes could be drawn less strongly; scale information could be added (the range of the vertical axes is about four hours, though precise values would be better read off a table of results); the level of α-blending might be varied across the display.

Figure 11 shows a special form of parallel coordinate plot. Usually each axis has its own scale and there is no natural ordering of the axes. Other examples of parallel coordinate plots can be found in other chapters of the Handbook.

5.3 Mosaic Plots

Mosaic plots display the counts in multivariate contingency tables. There are various types of mosaic plot (Hofmann; 2000) and a five-dimensional example of a doubledecker plot is displayed in Figure 12. The data are from a study of patterns of arrest based on 5226 cases in Toronto. Each column represents one combination of the four binary variables *Gender*, *Employed*, *Citzen*, and *Colour*. The width of a column is proportional to the number with that combination of factors. Those stopped who were not released later have been highlighted. Over 90% of those stopped were male. Some of the numbers of females in the possible eight combinations are too small to draw firm conclusions. Each pair of columns represents colour and the proportion not released amonst the males is lower amongst the whites for all combinations of other factors. The general decline in the level of highlighting across the male columns shows that the proportion not released is lower if the person is a citizen and lower still if they are employed. Figure 12 shows the difficulties in displaying data of this kind in a graphic for presentation. Colour, aspect ratio and size can make a big difference but labelling is the main problem.

5.4 Small Multiples and Trellis Displays

One way to avoid overloading a single large plot with information is to use a set of smaller, comparable plots instead. This can be effective for subgroup analyses (e.g., trellis displays for conditioning (Becker, Cleveland and Shyu; 1996)) or for geographic data (cf. micromaps Carr (2001)). A simple example is given in Figure 13. The boxplots on their own show that diesel cars

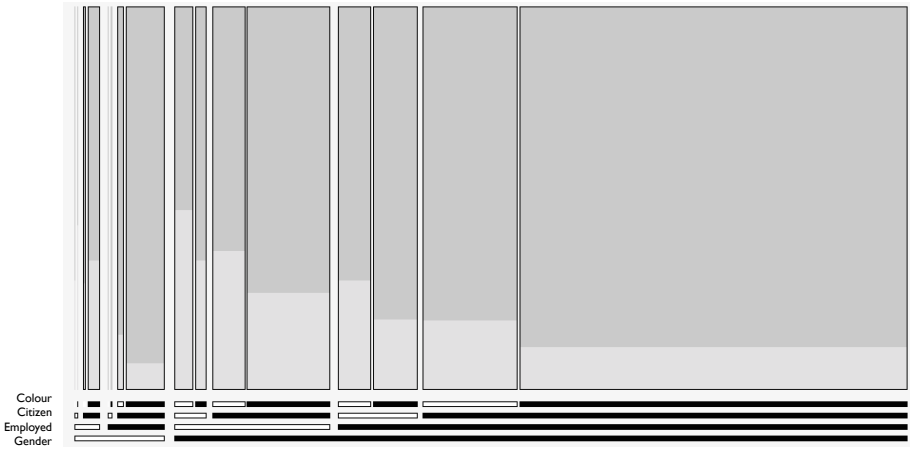

Fig. 12. A doubledecker plot of Toronto arrests data. Source: Fox (2003).

have generally lower fuel consumption (in Europe consumption is measured in litres/100km). The barchart on the left shows that little attention should be paid to the (Natural) Gas and Hybrid groups as few of these cars were measured. Should these two groups have been left out or perhaps replaced by dotplots? Small groups are always a problem. It should also be noted that the units for natural gas cars are different (kg/100km) from the others.

Small multiples can work well, but careful captioning is necessary to ensure that it is clear which smaller plot is which, and common scaling is obviously essential. Figure 14 is a trellis display of emissions' data for the 374 petrol or diesel cars. They have been grouped by engine type (the rows) and engine size (the columns). An equal count grouping has been used for engine size,

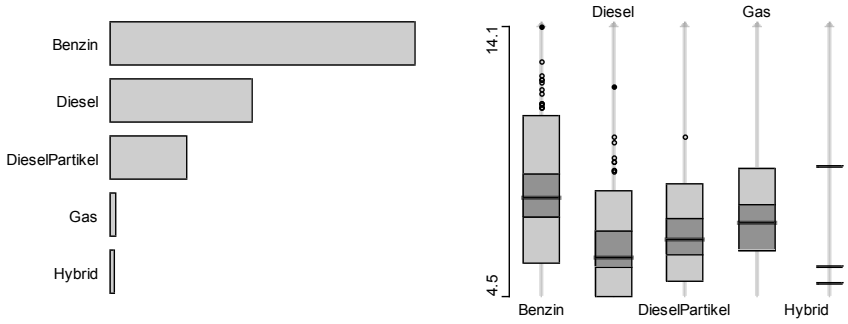

Fig. 13. Boxplots of fuel consumption by engine type data from Germany. The barchart shows the relative numbers of cars involved. The total number was 381. Source: http://www.adac.de, March 2006.

which is why the shaded parts of the *cc* bars have different lengths. Engine size seems to make little difference, as the plots in each row are similar to one another. The type of engine makes more difference, with diesel engines in particular being different from the other two types. There are a few local outliers amongst the petrol cars.

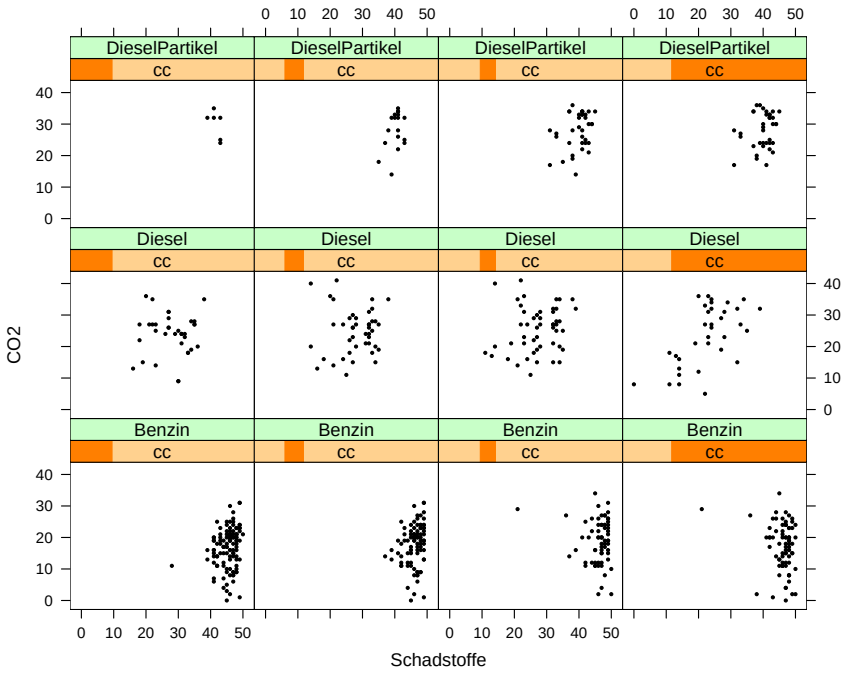

Fig. 14. A trellis display of car emissions data from Germany. Each panel is a scatterplot of two pollution measures. The rows refer to the type of engine and the columns to engine size. Source: http://www.adac.de, March 2006.

When several plots of the same kind are displayed, they can be plots of subsets of the same data, like in trellis displays, or plots of different variables for the same dataset, like in a parallel coordinates plot. It should always be obvious from the display which is the case.

5.5 Time Series and Maps

Time Series

Time series are special because of the strict ordering of the data, and good displays have to respect temporal ordering. It is useful to differentiate between

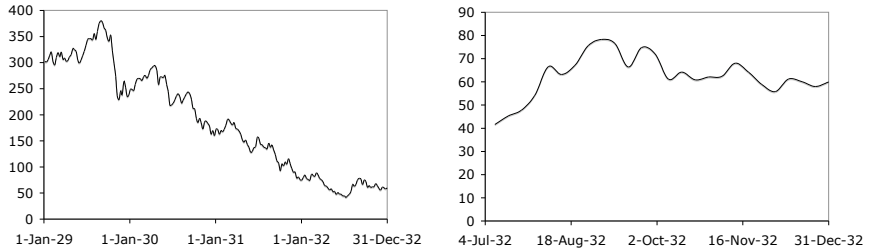

Fig. 15. Weekly Dow Jones Industrial Average: (a) Four years from 1929 to 1932. (b) Six months from July to December 1932. The maximum vertical axis value on the left is over four times the maximum on the right.

value measurements at particular time points (e.g. a patient's weight or a share price) and summary measurements over a period (e.g. how much the patient ate in the last month or how many shares were traded during the day).

Time scales have to be carefully chosen. The choice of time origin is particularly important, as anyone who looks at the advertised performance of financial funds will know. Time points for value measurements may not match the calendar scale (e.g.daily share prices only being available on days the market is open). Time units for summary measurements may be of unequal length (e.g. months). The time period chosen and the aspect ratio used for a time series plot can make a big difference to the interpretation of the data, see Figure 15.

If several time series are plotted in the same display, then it is necessary to ensure that they are properly aligned in time (e.g. two annual economic series may be published at different times of the year), that their vertical scales are matched (the common origin and the relative ranges) and that they can be distinguished from one another. Depending on the data, this can be tricky to do successfully.

Maps

Geographic data are complex to analyse, though graphical displays can be very informative. Bertin discussed a lot of ways of displaying geographic data in his book and there is a lot of sound advice in MacEachren's book (MacEachren; 1995), though more from a cartographic point of view. The main problems to be solved lie in the fact that areas do not reflect the relative importance of regions (e.g., Montana has fewer people than New York City but is much bigger) and spatial distance is not directly associated with similarity or nearness (e.g., where countries are divided by natural borders, like mountain ranges). There is a substantial research literature in geography on these and other display issues, such as how to use colour scales to show values (choropleth maps) and how to choose colour schemes (e.g., Colorbrewer referred to above). Some instructive examples can be found in the cancer atlas maps of the US Health

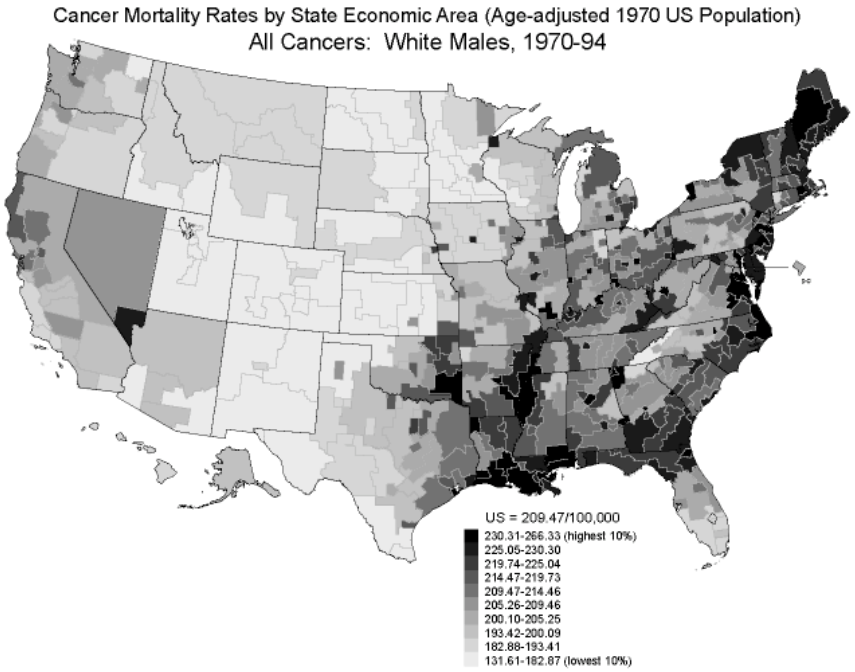

Fig. 16. Cancer mortality rates for white males in the US between 1970 and 1994 by State Economic Area. The scale has been chosen so that each interval contains 10% of the SEAs. Source: http://www3.cancer.gov/atlasplus/.

authorities on the web and in the book (Devesa, Grauman, Blot, Pennello, Hoover and Fraumeni; 1999). Figure 16 shows that cancer rates are highest along the East Coast and lowest in the Mid-West. State Economic Areas (SEAs) have been chosen, because using states oversmooths the data (consider Nevada in the West with its high cancer rate around Las Vegas, but its lower rate elsewhere), while using counties undersmooths. The map on the website is in colour, on a scale from deep red for high rates to dark blue for low. Naturally, this would not reproduce well in a gray-scale view, so the webpage provides the alternative version that is used here. Offering multiple versions of the same image on the web is readily possible but not often done. This is one of several reasons why the cancer atlas webpages are exemplary.

6 Practical Advice

6.1 Software

For a long time all graphics had to be prepared by draftsmen by hand. The volumes of the *Album de Statistique Graphique* produced towards the end of the

nineteenth century contain many exceptional displays which must have taken much painstaking preparation. Such graphics may be individually designed with special features for the particular data involved. Nowadays graphics are produced by software and this has tended to mean that certain default displays are adopted by many as a matter of course. If it takes a few minutes to prepare a graphic that is standard in your field, why bother to prepare something novel? This has advantages — standards avoid possible gross errors and are readily understood by readers familiar with them; and disadvantages — not all data fit the existing standards and interesting new information may be obscured rather than emphasised by a default display. As software becomes more sophisticated and user interfaces become more intuitive this may change. Currently (in 2006), there are software packages, which permit users substantial control over all aspects of the displays they wish to draw but these are still only for experts in the software (Murrell; 2005). It is reasonable to assume that there will be a steady progression to a situation where even non-experts can draw what they wish. Whether good graphics are the result will depend on the users' statistical good sense and on their design ability. The quality of a data visualization graphic depends on content and presentation just as the quality of a scientific article does. How has the quality of scientific articles changed since scientists have been able to prepare their own drafts with sophisticated text preparation software?

6.2 Bad Practice and Good Practice (Principles)

Sometimes it is easier to see what has gone wrong than to explain how to do something right. Take the simple task of preparing a barchart to display univariate categorical data. What could possibly go wrong? The bars may be too thin (or too fat); the gaps between the bars may be too narrow (or too wide): the labelling of the bars may be unclear (because it is difficult to fit long category names in); the order of the bars may be confusing; the vertical scale may be poorly chosen; there may be superfluous gridlines; irrelevant 3-D effects may have been used; colours or shading may have been unnecessarily added; the title may be misleading and the caption confusing. Doubtless there are even more ways of spoiling a barchart.

It is not possible to give rules to cover every eventuality. Guiding principles like the ones outlined in this chapter are needed.

7 And Finally

The lack of formal theory bedevils good graphics. The only way to make progress is through training in principles and through experience in practice. Paying attention to content, context and construction should ensure that sound and reliable graphics are produced. Adding design flair afterwards can add to the effect, so long as it is consistent with the aims of the graphic.

Gresham's Law in economics states that "Bad money drives out good." Fortunately this does not seem to apply to graphics, for while it is true that there are very many bad graphics displays prepared and published, there are also many very good ones. All serious data analysts and statisticians should strive for high standards of graphical display.

References

Becker, R., Cleveland, W. and Shyu, M.-J. (1996). The visual design and control of trellis display, *JCGS* **5**: 123–155.

Bertin, J. (1973). *Semiologie Graphique*, 2nd edn, Mouton-Gautier, The Hague.

Burn, D. (1993). Designing effective statistical graphs, *in* C. Rao (ed.), *Handbook of Statistics*, Vol. 9, Elsevier, pp. 745–773.

Carr, D. B. (2001). Designing linked micromap plots for states with many counties, *Statistics In Medicine* **20**: 1331–1339.

Cleveland, W. (1994). *The Elements of Graphing Data*, revised edn, Hobart Press, Summit, New Jersey, USA.

Dawson, R. (1995). The 'unusual episode' data revisited, *Journal of Statistics Education (Online)* **3**(3).

Devesa, S., Grauman, D., Blot, W., Pennello, G., Hoover, R. and Fraumeni, J. J. (1999). *Atlas of cancer mortality in the United States, 1950-1994*, US Govt Print Off, Washington, DC.

Everitt, B. (1993). *Cluster Analysis*, 3rd edn, Edward Arnold, London.

Fox, J. (2003). Effect displays in r for generalised linear models, *Journal of Statistical Software* **8**(15).

Hofmann, H. (2000). Exploring categorical data: interactive mosaic plots, *Metrika* **51**(1): 11–26.

Inselberg, A. (1999). Don't panic ... do it in parallel, *Computational Statistics* **14**(1): 53–77.

Izenman, A. and Sommer, C. (1988). Philatelic mixtures and multimodal densities, *Journal of the American Statistical Association* **83**(404): 941–953.

Kosslyn, S. (1994). *Elements of Graph Design*, Freeman, New York.

MacEachren, A. (1995). *How Maps Work*, Guildford Press, New York.

Murrell, P. (2005). *R Graphics*, Chapman & Hall, London.

Norman, D. (1988). *The Design of Everyday Things*, Doubleday, New York.

Playfair, W. (2005). *Playfair's Commercial and Political Atlas and Statistical Breviary*, Cambridge, London.

Putnam, R. (2000). *Bowling Alone*, Touchstone, New York.

Robbins, N. (2004). *Creating More Effective Graphs*, John Wiley.

Spence, R. (2001). *Information Visualization*, Adison-Wesley, New York.

Tufte, E. (2001). *The Visual Display of Quantitative Information*, 2nd edn, Graphic Press, Cheshire, Connecticut.

Wainer, H. (1997). *Visual Revelations*, Springer, New York.

Wainer, H. (2004). *Graphic Discovery: a Trout in the Milk and other Visual Adventures*, Princeton UP.

Wilkinson, L. (2005). *The Grammar of Graphics*, 2nd edn, Springer, New York.